**Goal:** Determine the DNA sequence of your DNA sample that was sent to the core and create a ‘virtual’ plasmid (e.g. YopH pNIC-Bsa4 or EhPTP pNIC-Bsa4).

1. Save this protocol using the following format Last Name\_First Name\_Analyze\_DNA\_Sequence in your student cloud (NOT TO THE COMPUTER).
2. Create a new word document and cut and paste your sequence that was emailed to you from the UT DNA sequencing facility.
3. Use screen shots to help demonstrate what you have done and include a short description. Replace the EXAMPLE: work with your own.
4. Your work will be assessed on clarity and thoroughness. Adding commentary between figures will help convey what you are doing.
5. When you have completed the protocol, you will then print your entire protocol (2 pgs per sheet) and glue into your notebook.
6. Email your final document to your instructor.

**Obtain a DNA Sequencing ‘read’ for the plasmid. This is a text file from the UT DNA Sequencing Facility that was sent to your email. To determine the sequence of a portion of our plasmid, the DNA was submitted to this core using two separate samples: one with a forward primer and one with a reverse primer. However, we can only have the sequence for the insert portion (target gene) but we know that the rest of the vector (the backbone) is from pNIC-Bsa4 (from the SGC – Structural Genomics Consortium). So, we should be able to determine the whole sequence if we know these parts and put them together.**

* Use ‘Tools’ ‘Word Count’ in Word to determine number of bases

**Enter # of bases here:** 1183

**NOTE: show each step** i.e. show sequence with N’s. Then show the sequence without N’s etc…..you will essentially have snapshots of each step.

EXAMPLE:

**dnaResults\_Plate\_21098\_Order\_140685\_Well\_3/EhPTP\_Forwardtxt**

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCATATGCACCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCCAATCCATGAAACTGCTGTTCGTTTGCCTGGGTAACATCTGCCGTTCTCCGGCTGCGGAAGCG

GTTATGAAAAAAGTTATCCAGAACCACCACCTGACCGAAAAATACATCTGTGACTCTGCGGGTACCTGCTCTTACCACGA

AGGTCAGCAGGCGGACTCTCGTATGCGTAAAGTTGGTAAATCTCGTGGTTACCAGGTTGACTCTATCTCTCGTCCGGTTG

TTTCTTCTGACTTCAAGAACTTTGACTACATCTTCGCGATGGACAACGACAACTACTACGAACTCCTGGACCGTTGCCCG

GAACAGTACAAACAGAAAATCTTCAAAATGGTAGACTTCTGCACCACCATCAAAACCACCGAAGTTCCGGACCCGTACTA

CGGTGGTGAAAAAGGTTTCCACCGTGTTATCGACATCCTGGAGGACGCGTGCGAAAACCTGATCATCAAACTGGAAGAAG

GTAAACTGATCAACTAACAGTAAAGGTGGATACGGATCCGAATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGC

ACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGANCAA

TAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGC

GAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGNGACCGCTACACTTGCCAG

CGCCCTANCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGNTNNNNNGCTTTCCCCGTCAAGCTCTAAATC

GGGGGCTCCNNTTNGGGNTNCNATTTANTGCTTTACGNCACNTCNANCNNNAAAAACTTGATTAGGGNGATNGNNNNNCG

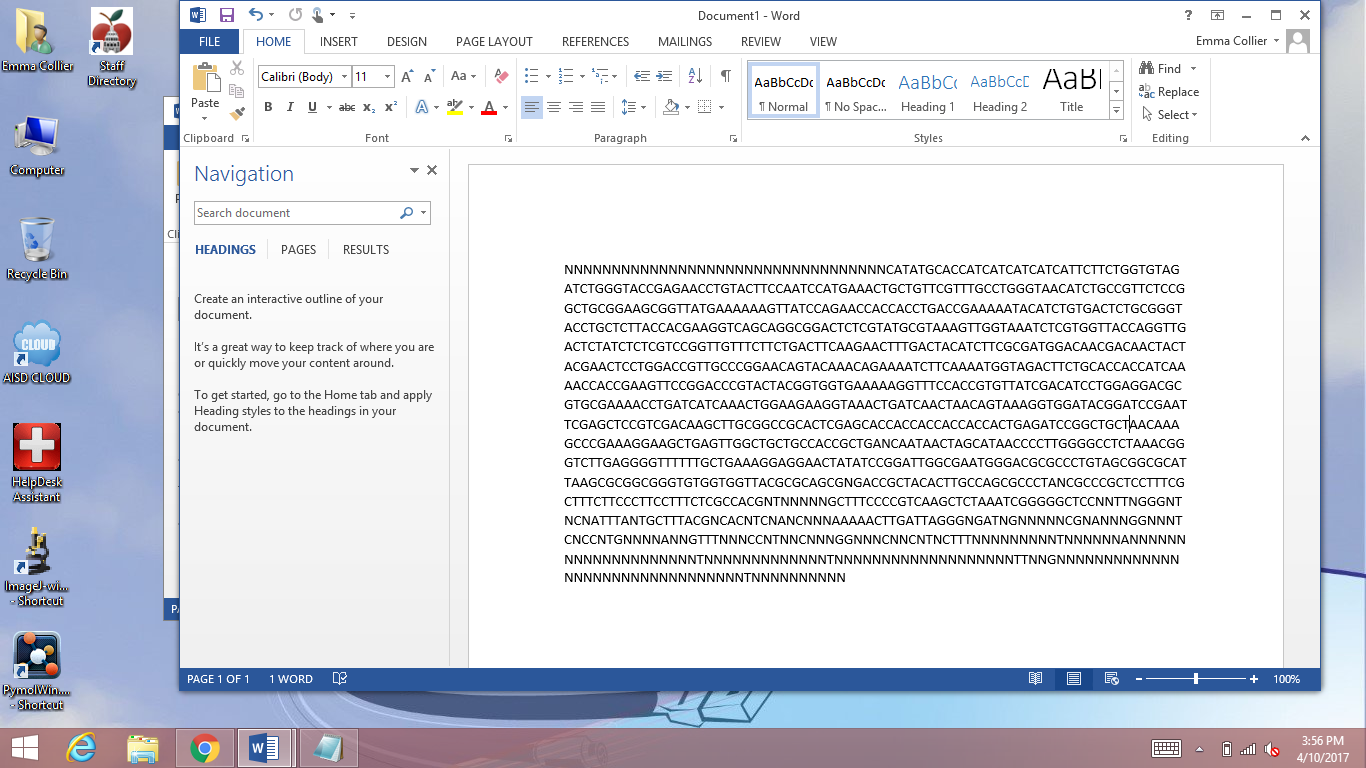
NANNNGGNNNTCNCCNTGNNNNANNGTTTNNNCCNTNNCNNNGGNNNCNNCNTNCTTTNNNNNNNNNTNNNNNNANNNNN

NNNNNNNNNNNNNNNTNNNNNNNNNNNNNTNNNNNNNNNNNNNNNNNNNTTNNGNNNNNNNNNNNNNNNNNNNNNNNNNN

NNNNNNTNNNNNNNNNN

**Remove the paragraph breaks**

* **When sequences are pasted into Word, they usually contain Paragraph breaks (¶) (you can see these by going to ‘Format > Reveal Formatting’ then selecting ‘Show all formatting marks’ at the bottom). This need to be removed because they will prevent searching for a specific set of bases in the sequence if it breaks over a line. So, use ‘Find and Replace’ to replace Paragraph marks with ‘nothing’ in the sequences for better searching and editing. Highlight the whole sequence, go to ‘Find and Replace’, select ‘More’ under Replace and then select ‘Special’ and ‘Paragraph Mark’. Leave ‘Replace with’ empty. Do ‘Replace All’ and then ‘No’ when it ask to search the rest of the document.**



EXAMPLE:

**EhPTP\_Forward.txt**

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCATATGCACCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCCAATCCATGAAACTGCTGTTCGTTTGCCTGGGTAACATCTGCCGTTCTCCGGCTGCGGAAGCGGTTATGAAAAAAGTTATCCAGAACCACCACCTGACCGAAAAATACATCTGTGACTCTGCGGGTACCTGCTCTTACCACGAAGGTCAGCAGGCGGACTCTCGTATGCGTAAAGTTGGTAAATCTCGTGGTTACCAGGTTGACTCTATCTCTCGTCCGGTTGTTTCTTCTGACTTCAAGAACTTTGACTACATCTTCGCGATGGACAACGACAACTACTACGAACTCCTGGACCGTTGCCCGGAACAGTACAAACAGAAAATCTTCAAAATGGTAGACTTCTGCACCACCATCAAAACCACCGAAGTTCCGGACCCGTACTACGGTGGTGAAAAAGGTTTCCACCGTGTTATCGACATCCTGGAGGACGCGTGCGAAAACCTGATCATCAAACTGGAAGAAGGTAAACTGATCAACTAACAGTAAAGGTGGATACGGATCCGAATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGANCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGNGACCGCTACACTTGCCAGCGCCCTANCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGNTNNNNNGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCNNTTNGGGNTNCNATTTANTGCTTTACGNCACNTCNANCNNNAAAAACTTGATTAGGGNGATNGNNNNNCGNANNNGGNNNTCNCCNTGNNNNANNGTTTNNNCCNTNNCNNNGGNNNCNNCNTNCTTTNNNNNNNNNTNNNNNNANNNNNNNNNNNNNNNNNNNNTNNNNNNNNNNNNNTNNNNNNNNNNNNNNNNNNNTTNNGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTNNNNNNNNNN

**These reads usually consist of about 1000 base pairs of sequence. There are, however, several bases that could not be determined by the sequencing machine. These are represented as ‘N’s’. You will need to remove many of these from the beginning and end to analyze the sequence better. Remove about 10-30 bp from the beginning and about 150-200 bp from the back end – this is kind of a judgment call on your part. For the backend – I recommend cutting back to the first instance of a string of 5 ‘N’s in a row.**

EXAMPLE:

**EhPTP\_Forward.txt - CHOPPED**

~~NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN~~CATATGCACCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCCAATCCATGAAACTGCTGTTCGTTTGCCTGGGTAACATCTGCCGTTCTCCGGCTGCGGAAGCGGTTATGAAAAAAGTTATCCAGAACCACCACCTGACCGAAAAATACATCTGTGACTCTGCGGGTACCTGCTCTTACCACGAAGGTCAGCAGGCGGACTCTCGTATGCGTAAAGTTGGTAAATCTCGTGGTTACCAGGTTGACTCTATCTCTCGTCCGGTTGTTTCTTCTGACTTCAAGAACTTTGACTACATCTTCGCGATGGACAACGACAACTACTACGAACTCCTGGACCGTTGCCCGGAACAGTACAAACAGAAAATCTTCAAAATGGTAGACTTCTGCACCACCATCAAAACCACCGAAGTTCCGGACCCGTACTACGGTGGTGAAAAAGGTTTCCACCGTGTTATCGACATCCTGGAGGACGCGTGCGAAAACCTGATCATCAAACTGGAAGAAGGTAAACTGATCAACTAACAGTAAAGGTGGATACGGATCCGAATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGANCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGNGACCGCTACACTTGCCAGCGCCCTANCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGNTNNNNNGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCNNTTNGGGNTNCNATTTANTGCTTTACGNCACNTCNANCNNNAAAAACTTGATTAGGGNGATNGNNNNNCGNANNNGGNNNTCNCCNTGNNNNANNGTTTNNNCCNTNNCNNNGGNNNCNNCNTNCTTT~~NNNNNNNNNTNNNNNNANNNNNNNNNNNNNNNNNNNNTNNNNNNNNNNNNNTNNNNNNNNNNNNNNNNNNNTTNNGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTNNNNNNNNNN~~

EXAMPLE:

**EhPTP\_Forward.txt - CHOPPED**

CATATGCACCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCCAATCCATGAAACTGCTGTTCGTTTGCCTGGGTAACATCTGCCGTTCTCCGGCTGCGGAAGCGGTTATGAAAAAAGTTATCCAGAACCACCACCTGACCGAAAAATACATCTGTGACTCTGCGGGTACCTGCTCTTACCACGAAGGTCAGCAGGCGGACTCTCGTATGCGTAAAGTTGGTAAATCTCGTGGTTACCAGGTTGACTCTATCTCTCGTCCGGTTGTTTCTTCTGACTTCAAGAACTTTGACTACATCTTCGCGATGGACAACGACAACTACTACGAACTCCTGGACCGTTGCCCGGAACAGTACAAACAGAAAATCTTCAAAATGGTAGACTTCTGCACCACCATCAAAACCACCGAAGTTCCGGACCCGTACTACGGTGGTGAAAAAGGTTTCCACCGTGTTATCGACATCCTGGAGGACGCGTGCGAAAACCTGATCATCAAACTGGAAGAAGGTAAACTGATCAACTAACAGTAAAGGTGGATACGGATCCGAATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGANCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGNGACCGCTACACTTGCCAGCGCCCTANCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGNTNNNNNGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCNNTTNGGGNTNCNATTTANTGCTTTACGNCACNTCNANCNNNAAAAACTTGATTAGGGNGATNGNNNNNCGNANNNGGNNNTCNCCNTGNNNNANNGTTTNNNCCNTNNCNNNGGNNNCNNCNTNCTTT

**Q**: **What are the consequences of cutting off too much?** You could accidently cut off a section of DNA that encodes the active site.

**TOO SHORT?**

**NOTE: if your DNA sequencing read from the core is particularly short, you may want to ‘stitch’ it together with the reverse sequence.**

**You would need to reverse complement the reverse and then figure out where it lines up with the forward. Paste them both here – one above the other. Then join them together at that point to make a fuller sequence of your gene.**

* + - <http://www.bioinformatics.org/sms2/rev_comp.html>

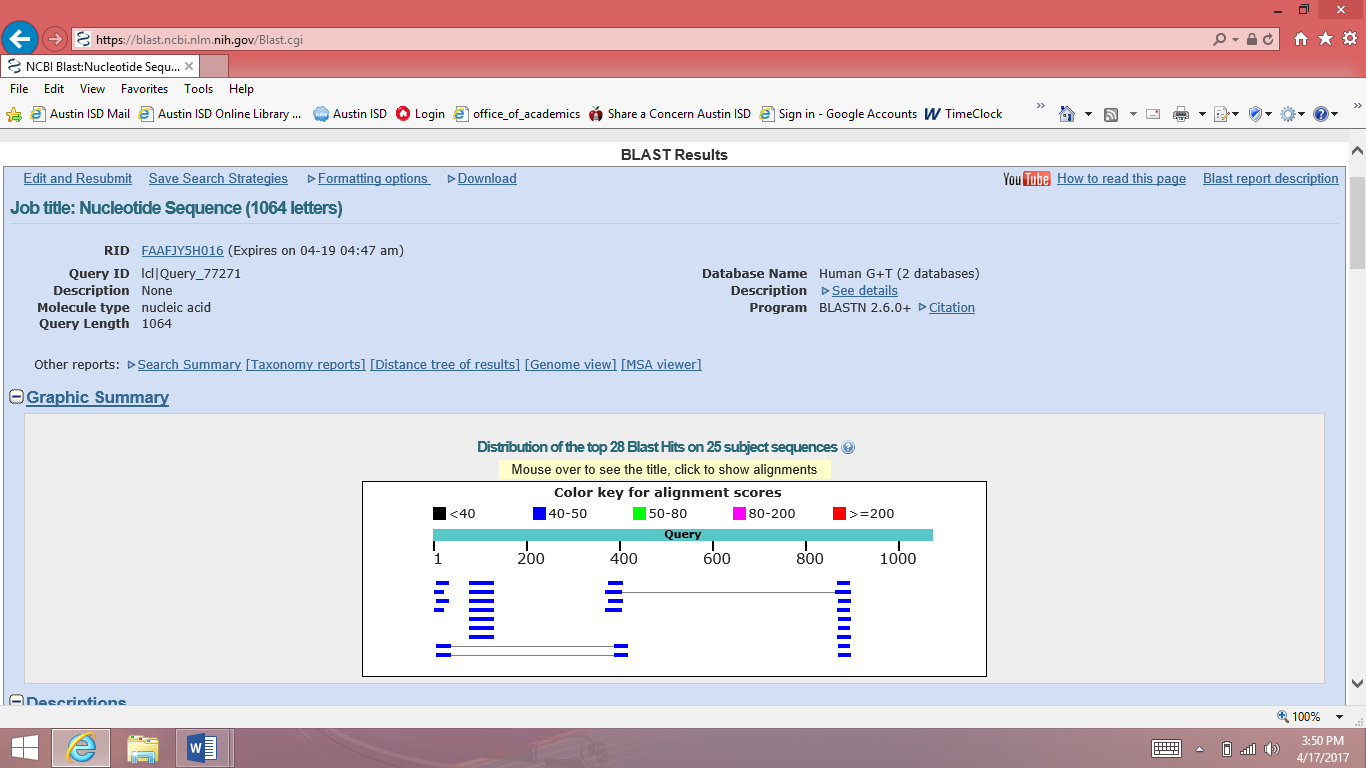
Do a BLAST by going to the following site: <http://www.ncbi.nlm.nih.gov/>

Choose BLAST under “Popular Resources” on the right side of the screen.

**Q. What is a BLAST?**

1. Choose ‘nucleotide blast’ under WEB BLAST
2. Cut and paste the sequence of your concatenated DNA sequencing read from DNA core
3. Choose ‘Human genomic plus transcript (Human G+T)’ under ‘Database’
4. Select ‘Somewhat similar sequences (blastn)’
5. Begin Search by selecting “BLAST”

**Show a screenshot of the top part of the results page with the color bar.**

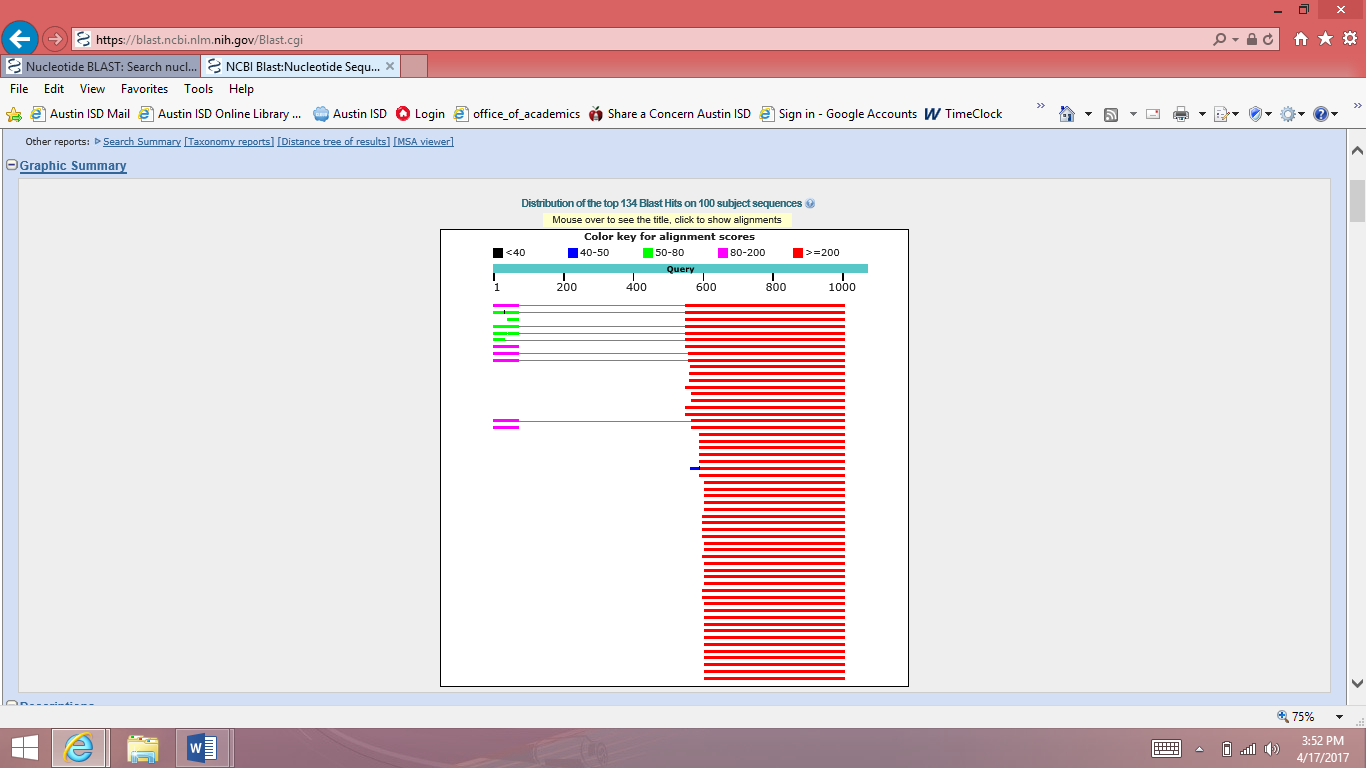


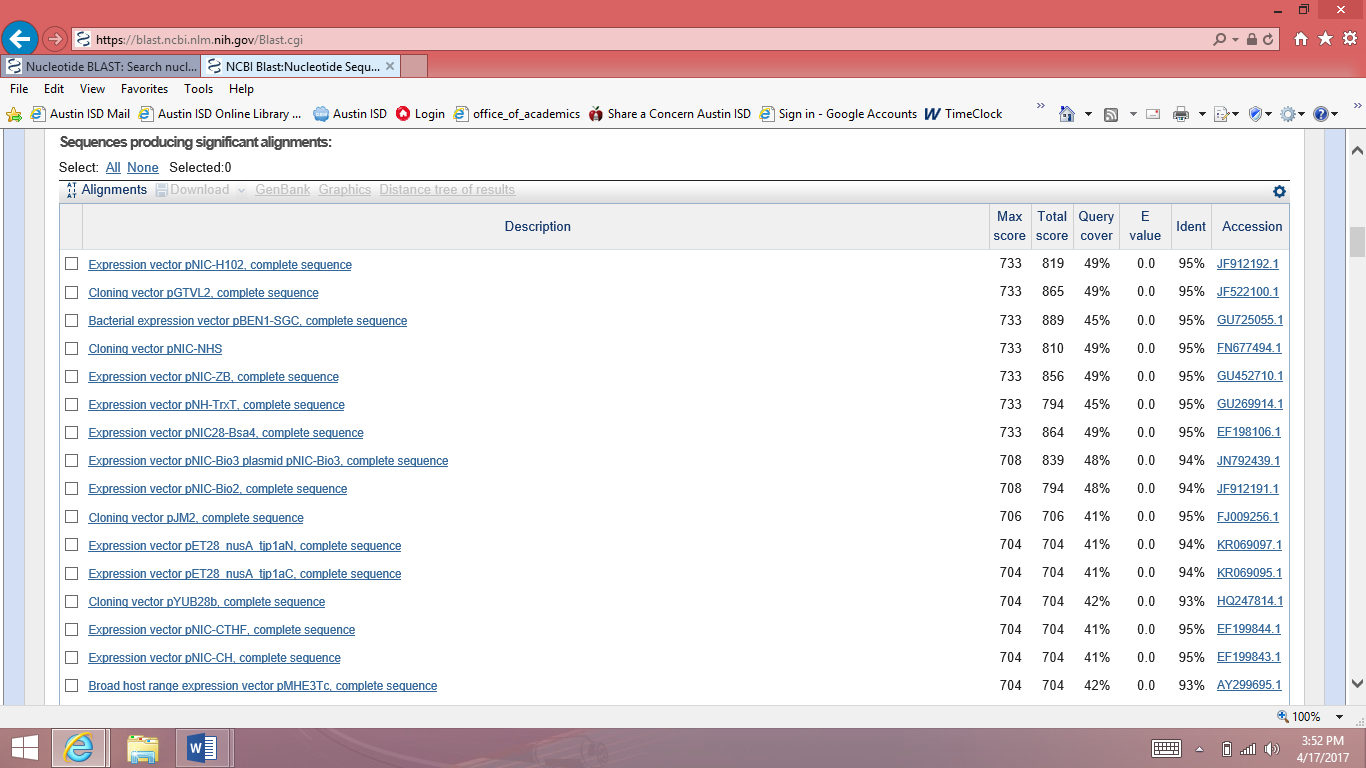
**Q: Is your sequence in the human genome?** Yes, and it should be.

(Should it be?)

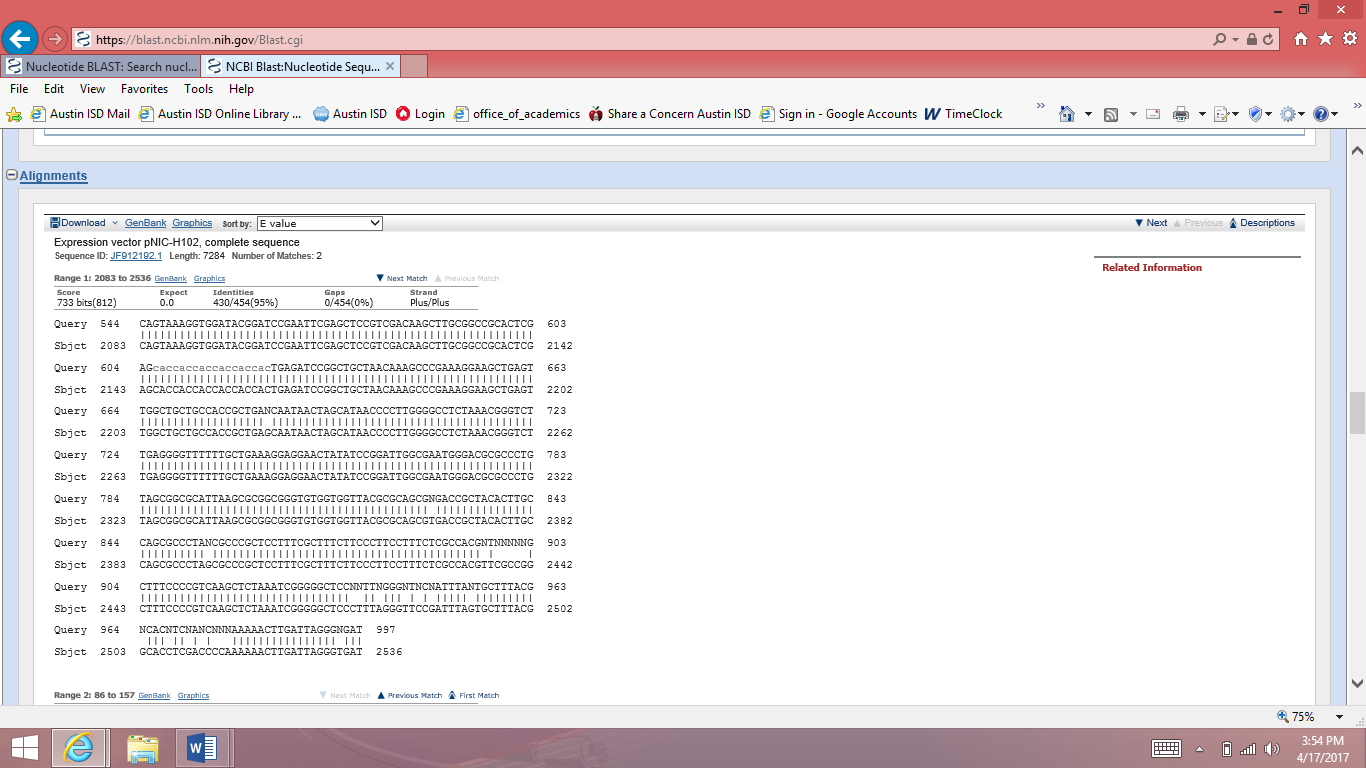
1. Do another BLAST to see if your sequence is found in other genomes.
2. Under ‘Database’ choose ‘Nucleotide collection (nr/nt)’ from the drop down options
   1. This database represents sequences from ALL organisms
3. Select ‘Somewhat similar sequences (blastn)’
4. Begin Search by selecting “BLAST”

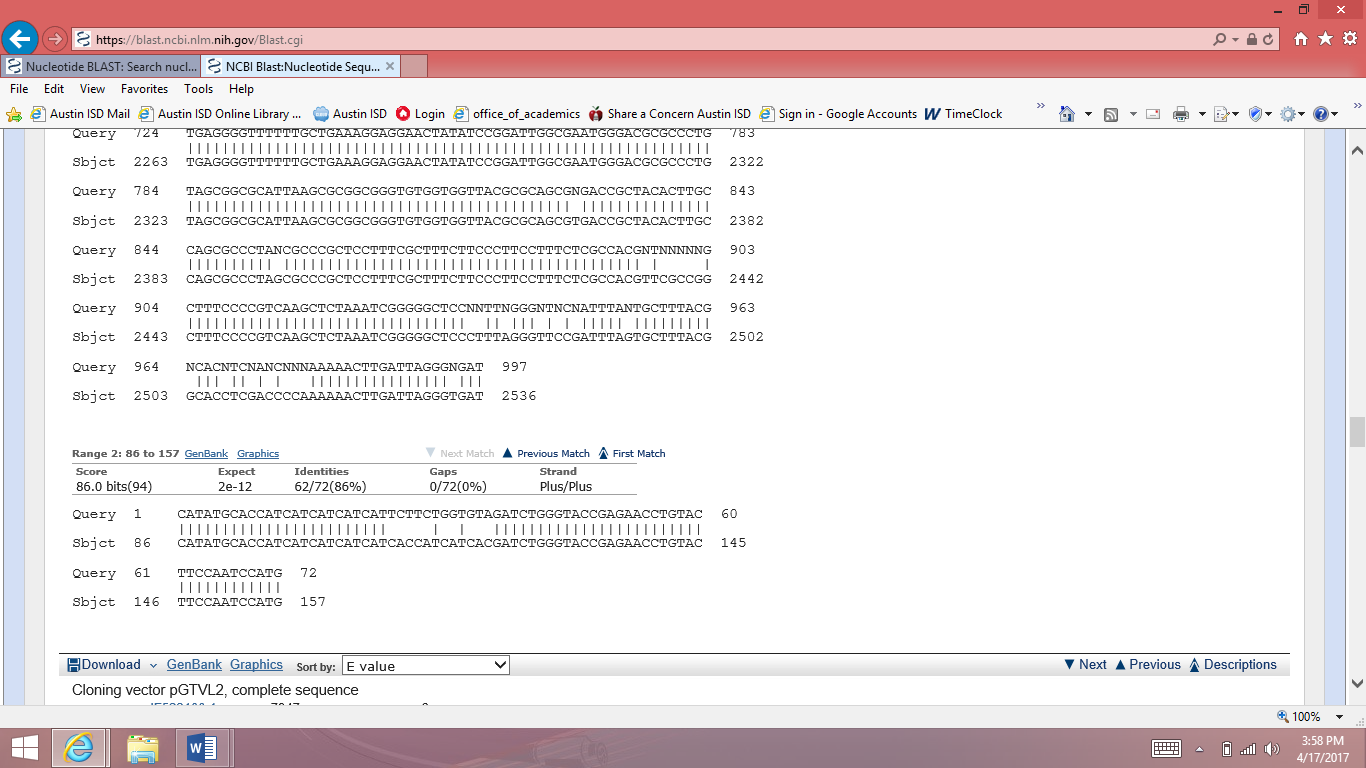
**Show a screenshot of the top part of the results page with the color bar and about 5-10 top hits (with the Query Coverage and Idents shown)**





**Show a screenshot of the Pairwise Alignment of the top hit**. You can find this by clicking on the name of the top hit – it will then link you to the bottom of the page to show the alignment with the two sequence on top of one another. (the Query and the Subject)





On the BLAST results under Descriptions,

* **Query coverage** represents the proportion of the length of the DNA sequence which the server was able to find some matching regions.
* **Identities** represents the percentage of bases within the query coverage that were matching nucleotides between the two sequences queried (e.g. Adenine (A) with Adenine (A) or G to G, or C to C, or T to T)

**Q: Did you find anything that is similar to your sequence?** Yes, the expression vector pNIC-H102 (complete sequence).

**Q: How close are they?** Nearly the entire expression vector is contained in the DNA, so identical in some segments.

**Query coverage?** 49%

**Identities?** 95%

**Q: What gene is it?** SacB

**Q: Is your sequence in any other genome?** Humans, plants, bacteria, and archaea

**Comparison to Known CDS (a.k.a. - Pairwise Alignment):**

1. To get the mRNA sequence for your target gene from the NCBI database using the RefSeq identifier: <https://www.ncbi.nlm.nih.gov/>
2. Paste in either of these sequence identifiers:

**NCBI RefSeq for YopH**: NC\_004836.1

(note that full length YopH is 467 aa but our portion is about 306 aa)

**OR**

**NCBI RefSeq for EhPTP**: XM\_651267.1

1. Then go to ‘Gene’

For EhPTP – click on the 1 entry in Gene

For YopH – search the 4 pages until you find ‘YopH’

1. Copy the FASTA Sequence (Right above the colored bars – click on ‘FASTA’)
2. Enter this into the QUERY field of NCBI BLASTN

<https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch>

1. Make sure you are doing blastn not blastp at the top
2. Check the ‘Align two or more sequences’
3. Input the read from the DNA Sequencing core in the bottom (the SBJCT)
4. Select ‘Highly Similar Sequences’, Hit ‘BLAST’
5. See results to determine if the sequences are similar.

Can you see any mismatches between the two? If so, is it a problem? Why or why not?

**If you find anything – highlight it in your pairwise alignment**

Note: widen your margins when pasting to WORD. So, that the formatting will be readable – the bars should line up with the letters. You can also reduce the font size.

It turns out that these **SHOULD NOT MATCH**! This mismatching is due to the way we clone our genes. Since the expression organism (*E. coli*) is different than the original organism (e.g. *Yersinia pestis or Entamoeba histolytica*) we often want to codon-optimize the DNA sequence so that it will express the protein to a higher yield in the *E. coli* cells. Codon optimization essentially changes the DNA bases around. So, to determine if our DNA sequence from the core really is our target – we should translate the sequence into protein because the protein stays the same, even when we swap out codons in the DNA. It will be easier then to compare protein to protein.

**Lost in Translation:**

Translate the original sequence into protein, but first remove all the N’s from the original sequence by finding and replacing as you did above.

Go to Sequence Manipulation Suite

<http://www.bioinformatics.org/sms2/>

You can translate your DNA sequence into protein using ‘Translate’

Use the standard genetic code

Try all three reading frames (1,2,3)

**Translation for Reading Frame 1:**

HMHHHHHHSSGVDLGTENLYFQSMKLLFVCLGNICRSPAAEAVMKKVIQNHHLTEKYICDSAGTCSYHEGQQADSRMRKVGKSRGYQVDSISRPVVSSDFKNFDYIFAMDNDNYYELLDRCPEQYKQKIFKMVDFCTTIKTTEVPDPYYGGEKGFHRVIDILEDACENLIIKLEEGKLINQRWIRIRIRAPSTSLRPHSSTTTTTTEIRLLTKPERKLSWLLPPLTITSITPWGLTGLEGFFAERRNYIRIGEWDAPCSGALSAAGVVVTRSGPLHLPAPYARSFRFLPFLSRHVLSPSSSKSGAPWVIYALRTHKNLIRGCEVPEFPRPL

Translation for Reading Frame 2:

ICTIIIIILLV\*IWVPRTCTSNP\*NCCSFAWVTSAVLRLRKRL\*KKLSRTTT\*PKNTSVT

LRVPALTTKVSRRTLVCVKLVNLVVTRLTLSLVRLFLLTSRTLTTSSRWTTTTTTNSWTV

ARNSTNRKSSKW\*TSAPPSKPPKFRTRTTVVKKVSTVLSTSWRTRAKT\*SSNWKKVN\*ST

NSKGGYGSEFELRRQACGRTRAPPPPPLRSGC\*QSPKGS\*VGCCHR\*Q\*LA\*PLGASKRV

LRGFLLKGGTISGLANGTRPVAAH\*ARRVWWLRAADRYTCQRPTPAPFAFFPSFLATCFP

RQALNRGLLGSFMLYALTKT\*LGDARSLSFLGLF

Translation for Reading Frame 3:

YAPSSSSFFWCRSGYREPVLPIHETAVRLPG\*HLPFSGCGSGYEKSYPEPPPDRKIHL\*L

CGYLLLPRRSAGGLSYA\*SW\*ISWLPG\*LYLSSGCFF\*LQEL\*LHLRDGQRQLLRTPGPL

PGTVQTENLQNGRLLHHHQNHRSSGPVLRW\*KRFPPCYRHPGGRVRKPDHQTGRR\*TDQL

TVKVDTDPNSSSVDKLAAALEHHHHHH\*DPAANKARKEAELAAATADNN\*HNPLGPLNGS

\*GVFC\*KEELYPDWRMGRAL\*RRIKRGGCGGYAQRTATLASALRPLLSLSSLPFSPRAFP

VKL\*IGGSLGHLCFTHSQKLD\*GMRGP\*VSSAS

Which one do you think is right?

Can you find the starting amino acid (M)? Can you find the His6 tag (6 H’s in a row)?

Highlight them in WORD

The Start amino acid can be in green: M

The His6 Tag can be in Yellow

**After you have determined which reading frame you should use remove all of the \* from the sequence using find and replace.**

**Compare Protein Translations:**

After translating your sequence to protein grab the known protein sequence:

**YopH**

PDB entry: 2Y2F

Go to Display Files >> Fasta

MRERPHTSGHHGAGEARATAPSTVSPYGPEARAELSSRLTTLRNTLAPATNDPRYLQACG

GEKLNRFRDIQCRRQTAVRADLNANYIQVGNTRTIACQYPLQSQLESHFRMLAENRTPVL

AVLASSSEIANQRFGMPDYFRQSGTYGSITVESKMTQQVGLGDGIMADMYTLTIREAGQK

TISVPVVHVGNWPDQTAVSSEVTKALASLVDQTAETKRNMYESKGSSAVADDSKLRPVIH

CRAGVGRTAQLIGAMCMNDSRNSQLSVEDMVSQMRVQRNGIMVQKDEQLDVLIKLAEGQG

RPLLNS

Or this one is close (one amino acid difference) WP\_068794378.1

**EhPTP**

PDB entry: 3JS5

Go to Display Files >> Fasta

>3JS5:A|PDBID|CHAIN|SEQUENCE

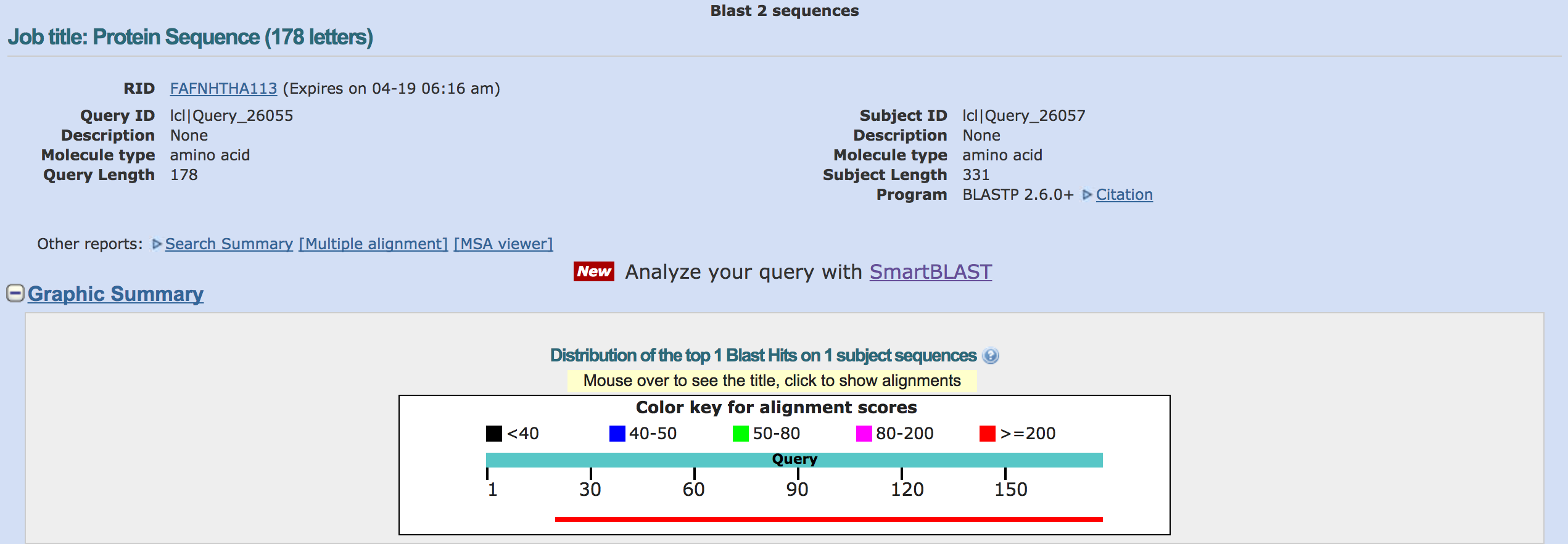
MAHHHHHHMGTLEAQTQGPGSMKLLFVCLGNICRSPAAEAVMKKVIQNHHLTEKYICDSAGTCSYHEGQQADSRMRKVGKSRGYQVDSISRPVVSSDFKNFDYIFAMDNDNYYELLDRCPEQYKQKIFKMVDFCTTIKTTEVPDPYYGGEKGFHRVIDILEDACENLIIKLEEGKLIN

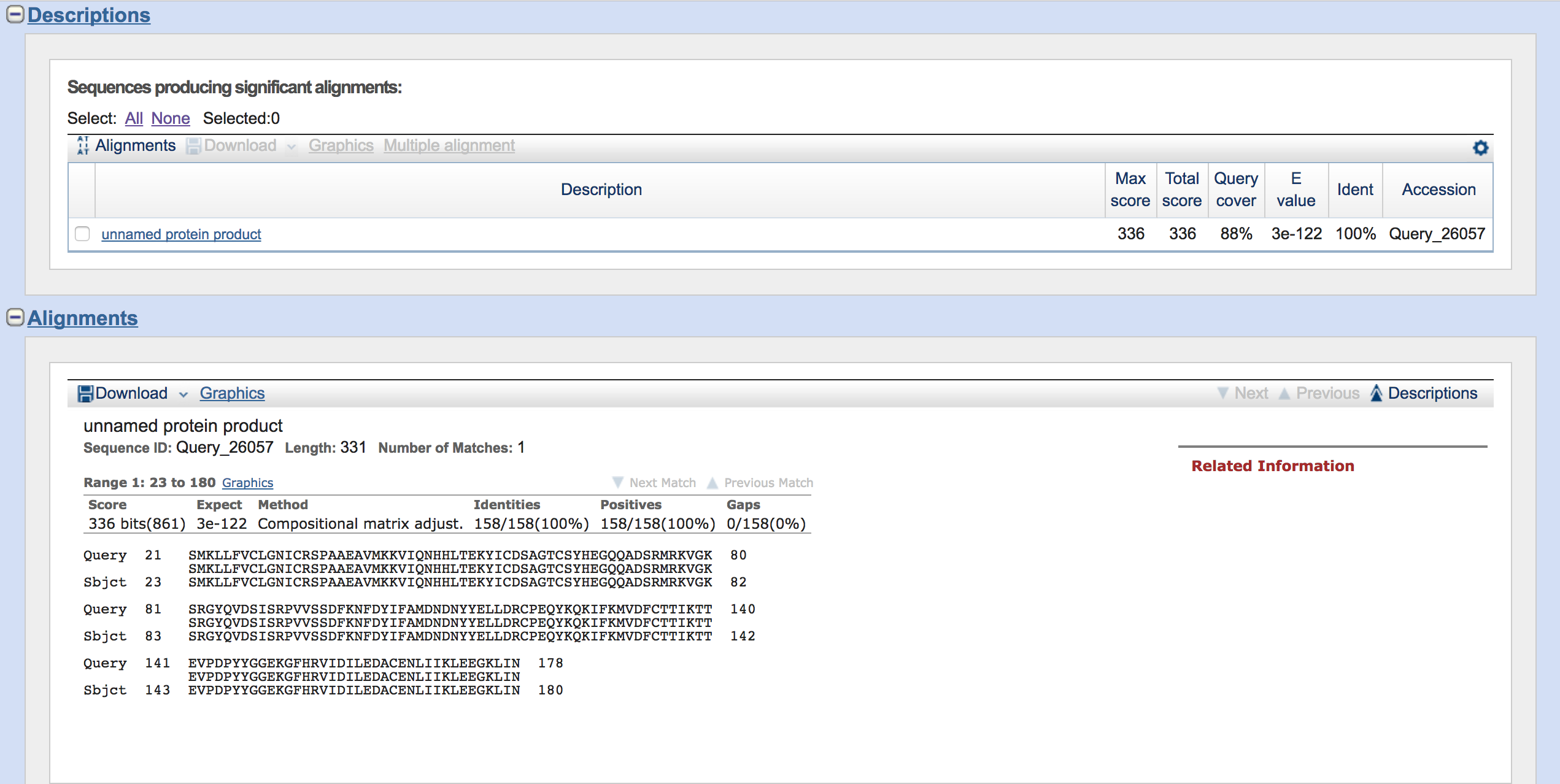
Same as: XP\_656359.1

Or this one is close (one amino acid difference) XP\_653357.1

* Go to BLAST and go to BLASTP for protein (just a tab on the BLASTN page).
* Check the box to ‘Align two or more sequences’
* paste both into BLASTP to check that any DNA mutations are not a problem for the protein.

Paste your results in to this document.





**Identities, Mutations, & Conservative Changes**

When comparing sequences, there can be missing residues (amino acids), changes for one residue for another or insertions. Identical residues are shown as being the same letter. Conservative changes with a ‘+’ plus sign (e.g. Threonine (T) and Serine (S) at position 255 in the example below).This is a conservative change because the amino acid is different between the two, but they share similar properties. Both Threonine and Serine contain an hydroxyl (-OH) group which make them polar but uncharged. They are also roughly the same size. So, a mutation from Threonine to Serine may have a minimal impact on the protein at that location => conservative change. In contrast, point mutations are shown as blanks spaces (e.g. Threonine (T) to Histidine (H) mutation at position 253).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Score** | **Expect** | **Method** | **Identities** | **Positives** | **Gaps** |
| 336 bits(861) | 3e-122 | Compositional matrix adjust. | 158/158(100%) | 158/158(100%) | 0/158(0%) |

Query 21 SMKLLFVCLGNICRSPAAEAVMKKVIQNHHLTEKYICDSAGTCSYHEGQQADSRMRKVGK 80

SMKLLFVCLGNICRSPAAEAVMKKVIQNHHLTEKYICDSAGTCSYHEGQQADSRMRKVGK

Sbjct 23 SMKLLFVCLGNICRSPAAEAVMKKVIQNHHLTEKYICDSAGTCSYHEGQQADSRMRKVGK 82

Query 81 SRGYQVDSISRPVVSSDFKNFDYIFAMDNDNYYELLDRCPEQYKQKIFKMVDFCTTIKTT 140

SRGYQVDSISRPVVSSDFKNFDYIFAMDNDNYYELLDRCPEQYKQKIFKMVDFCTTIKTT

Sbjct 83 SRGYQVDSISRPVVSSDFKNFDYIFAMDNDNYYELLDRCPEQYKQKIFKMVDFCTTIKTT 142

Query 141 EVPDPYYGGEKGFHRVIDILEDACENLIIKLEEGKLIN 178

EVPDPYYGGEKGFHRVIDILEDACENLIIKLEEGKLIN

Sbjct 143 EVPDPYYGGEKGFHRVIDILEDACENLIIKLEEGKLIN 180

**Q: Can there be ‘conservative’ changes in DNA sequences?** Yes

**Insertion into Vector backbone:**

Obtain pNIC-Bsa4 sequence from Google Drive.

<https://drive.google.com/open?id=1nut0QlGpAoubVtVZI1LXpn2ZkKUSHOXUvc2Ev3oSrNY>

Note on terminology: backbone = vector = plasmid = pNIC-Bsa4 = DNA

We need to filter this to get it in a better format.

* Go to Sequence Manipulation Suite: <http://www.bioinformatics.org/sms2/>
* Choose Filter DNA on the left
* Paste the vector sequence along with all the numbers. Use default options.
* Copy the filtered sequence to Word.
* Then remove any Paragraph breaks as you have done before.

Now find out where the sequence of your DNA sequencing Read from the DNA core matches up with the pNIC-Bsa4 backbone sequence

* Open up the pNIC-Bsa4 sequence
* We will then highlight in different colors the different feature of the gene insert and the backbone vector so that we can ‘map’ them and figure out where the gene should go in the plasmid vector.

In your pNIC-Bsa4 sequence, use Ctrl+F to FIND the pLIC-forward primer site.

* Highlight it in gray
* pLIC-for primer site TGTGAGCGGATAACAATTCC

Do the same for the pLIC-rev but highlight it in pink

* pLIC-rev primer site AGCAGCCAACTCAGCTTCC
  + NOTE: you must reverse complement this first to find it on the sense strand that we are using for pNIC-Bsa4 (remember the structure of DNA is a double helix: a sense strand and an anti-sense strand)
    - <http://www.bioinformatics.org/sms2/rev_comp.html>
    - Reverse complement: GGAAGCTGAGTTGGCTGCT

NOTE: if you cannot find the pLIC-rev in your gene sequence – then try the CLONING primers – because they delineate exactly where the gene ends.

Primers for LIC cloning:

Upstream: add TACTTCCAATCCATG to the 5’ end (ATG in-frame with the desired coding sequence).

Downstream: add TATCCACCTTTACTG to 5’ end of downstream primer; add termination codon, if necessary.

Find the T7 promoter sequence and mark it in blue: TAATACGACTCACTATAGG

Your gene sequence will fit within these regions.

Now for both the gene sequence and the backbone vector of the plasmid (pNIC-Bsa4)

Find and highlight the Start codon in green: ATG

This is the 3-letter DNA code that denotes the beginning of the protein

The Stop codon in red: TAA

* This is the 3-lletter DNA code that denotes the end of the protein
* Note that there will be many TAA’s but the STOP codon TAA has to be ‘in-frame’ with the ATG start codon. That is the distance between them must be in multiples of three (3) since each codon contains 3 bases per residue.
* **Q: Which one do you think is right?** The TAA that is 2,037bp away from the start codon.
* Maybe proceed through the next steps and come back to this question.
* Use the translate feature in the DNA Manipulation Suite to figure out if you are making the whole protein for a given STOP codon.

The 6xHIS tag: in YELLOW (for this you need to figure out what codons make HISTIDINE)

* Search for several of these in a row
* This should show up in BOTH sequences

**Open the pNIC-Bsa4 sequence and insert the section of the DNA sequencing read**

**Underline the DNA Sequencing read sequence once it is in the pNIC-Bsa4 sequence**

**Now copy and paste the highlighted, underline version of pNIC-Bsa4 below**

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGCACCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCCAATCCATGGAGACCGACGTCCACATATACCTGCCGTTCACTATTATTTAGTGAAATGAGATATTATGATATTTTCTGAATTGTGATTAAAAAGGCAACTTTATGCCCATGCAACAGAAACTATAAAAAATACAGAGAATGAAAAGAAACAGATAGATTTTTTAGTTCTTTAGGCCCGTAGTCTGCAAATCCTTTTATGATTTTCTATCAAACAAAAGAGGAAAATAGACCAGTTGCAATCCAAACGAGAGTCTAATAGAATGAGGTCGAAAAGTAAATCGCGCGGGTTTGTTACTGATAAAGCAGGCAAGACCTAAAATGTGTAAAGGGCAAAGTGTATACTTTGGCGTCACCCCTTACATATTTTAGGTCTTTTTTTATTGTGCGTAACTAACTTGCCATCTTCAAACAGGAGGGCTGGAAGAAGCAGACCGCTAACACAGTACATAAAAAAGGAGACATGAACGATGCACCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCCAATCCATGAAACTGCTGTTCGTTTGCCTGGGTAACATCTGCCGTTCTCCGGCTGCGGAAGCGGTTATGAAAAAAGTTATCCAGAACCACCACCTGACCGAAAAATACATCTGTGACTCTGCGGGTACCTGCTCTTACCACGAAGGTCAGCAGGCGGACTCTCGTATGCGTAAAGTTGGTAAATCTCGTGGTTACCAGGTTGACTCTATCTCTCGTCCGGTTGTTTCTTCTGACTTCAAGAACTTTGACTACATCTTCGCGATGGACAACGACAACTACTACGAACTCCTGGACCGTTGCCCGGAACAGTACAAACAGAAAATCTTCAAAATGGTAGACTTCTGCACCACCATCAAAACCACCGAAGTTCCGGACCCGTACTACGGTGGTGAAAAAGGTTTCCACCGTGTTATCGACATCCTGGAGGACGCGTGCGAAAACCTGATCATCAAACTGGAAGAAGGTAAACTGATCAACTAAAAACGCAAAAGAAAATGCCGATATCCTATTGGCATTGACGGTCTCCAGTAAAGGTGGATACGGATCCGAATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGTTTACAATTTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATCGAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCATGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGGCATAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTTTCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCGACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAAT

**Then copy and paste your target gene sequence from the HIS tag all the way to the pLIC-Rev primer site.**

**Then paste this into the pNIC-Bsa4 sequence to replace what was already there.**

**Congratulations! -** Now, you have the complete sequence for **YopH-pNIC-Bsa4** or **EhPTP-pNIC-Bsa4!**

Save your sequence because we will use it to design Site Directed Mutagenesis Primers

**BONUS Questions:**

1. **What is the gene that is already in pNIC-Bsa4 that your gene replaces?**

It replaces the SCAF11 (SR-Related CTD Associated Factor 11) protein coding gene.

1. **What Restriction Enzymes will cut your plasmid?**

DNA Polymerases